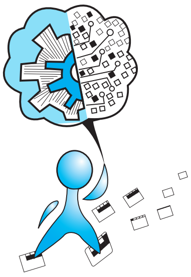

Data Quality and Machine Learning: What's the Relationship?

Jitendra Yasaswi K.

B.Tech IT, 2014. Research Scientist at the Robert Bosch Research and Technology Center, Bangalore.
Email: katta.jitendrayasaswibharadwaj@in.bosch.com

The buzzword AI is an increasingly used term in today's technological world. In the last decade, we have witnessed enormous growth in AI enabled applications like never before. At the start of last decade, the focus of Machine Learning (ML) applications was to tag or classify images and today we have ML algorithms that can even generate images (e.g. DALL-E 2). Similarly, a decade ago we started with applications that can predict sentiment of a user from movie reviews and today we have systems that can generate text and even fill parts of software code (e.g. Github Copilot).

Many ML practitioners consider the year 2012 as the breakthrough year for AI where they started to believe that deep learning models indeed work better than traditional ML methods. This success can be attributed to two important factors 1) the availability of large training datasets and 2) the availability of infrastructure to train large deep learning models. If you look at any ML problem setting in the real world, it has two main parts which are the *core model* part and the *core data* part. The core model part deals with designing models (say deep learning models), optimizers, coming up with suitable training schemes, validating and maintaining these trained models. As a result of this, we have witnessed a number of important technical developments on the *core model* part like improvements in Convolutional Neural Networks (CNNs), Variational AutoEncoders (VAEs), Generative Adversarial Networks (GANs) and Transformer Models. On the other hand, talking about the *core data* aspect, apart from collecting huge datasets, very little or no emphasis is laid on the quality of the training data and managing it.

Many challenges that AI based applications face in the real world are due to imperfections in data. Imagine a scenario where you are developing an AI algorithm which is to be deployed in a real world application. To ensure that your AI algorithm works with extreme reliability, this algorithm has to be trained with data that contains even rarest of the rare situations that the application might encounter in the real world. Recently, Tesla's autopilot feature had mistaken a horse carriage for a truck and this could be due to the long tailedness of the data (the ML model(s) deployed in autopilot may have encountered very few or no scenes containing horse carriage during training). This is why "*Data-centric AI*" is needed. Here, the focus is on collecting good quality training data and ensuring good quality labelling. This also includes ensuring good data coverage, which could be done through search, retrieval of rare instances, usage of data augmentations (self-supervision) and synthetic data generation to generate rare scenarios or to fill domain gaps. This also includes methods to identify and handle data drifts post model deployment.

There is a saying that "*Your model is as good as your data*" and good quality data is a must to build any successful ML pipeline. As articulated by Andrew Ng in "MLOps: From Model-centric to Data-centric AI", the future is moving towards Data-centric AI.

About the Author

Jitendra Yasaswi Katta is a Research Scientist at the Robert Bosch Research and Technology Center in Bangalore, India. His work spans the intersection of Computer Vision and Machine Learning, with special focus on self-supervised learning and representation learning. Prior to this, he worked at Teradata Labs where he was part of the R&D team developing Machine Learning applications for prediction tasks. He received his Master's degree (MS by Research) in Computer Science and Engineering from IIIT Hyderabad and Bachelor's degree in Information Technology from JNTUK UCE Vizianagaram.